DOCUMENT RESUME

ED 393 880 TM 024 756

AUTHOR Schwarz, Julie A.; Collins, Michelle L.
TITLE Improving the Reliability of a Direct Writing Skills Assessment.
PUB DATE Jun 95
NOTE 31p.; Paper presented at the Annual Meeting of the International Personnel Management Association Assessment Council Conference (19th, New Orleans, LA, June 25-29, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Assessment; *Graduate Students; Graduate Study; Higher Education; *Interrater Reliability; *Police; Rating Scales; *Scoring; Test Construction; *Writing Tests
IDENTIFIERS *Behaviorally Anchored Rating Scales; Direct Writing Assessment; *Frame of Reference Model

ABSTRACT

Behaviorally Anchored Rating Scales (BARS) were developed to score responses from a previously designed police written communication test that lacked reliability. Rating scales for each of the 9 dimensions of the test consisted of the scale definition and a 5-point continuum, with the scores of 5, 3, and 1 defined by specified behavioral incidents. Ten raters (graduate students) were trained to score the test using a Frame-of-Reference technique. Fifty tests previously completed by police incumbents were re-scored. Each test was scored by two raters in order to assess interrater reliabilities. Interrater reliabilities and internal consistency were found to improve. An appendix contains definitions of behaviorally based anchors. (Contains 2 tables and 38 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JULIE SCHWARZ

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 393 880

Improving the Reliability of a Direct Writing Skills Assessment

Julie A. Schwarz

University of North Texas

Michelle L. Collins

Personnel Decisions, Inc.

Paper presented at the Nineteenth Annual IPMAAC Conference.

BEST COPY AVAILABLE

TM024756

ABSTRACT

Behaviorally Anchored Rating Scales (BARS) were developed to score responses from a previously designed police written communication test which lacked reliability. Ten raters were trained to score the test using a Frame-of-Reference (FOR) technique. Fifty tests previously completed by police incumbents were re-scored. Each test was scored by two raters in order to assess interrater reliabilities. Interrater reliabilities and internal consistency were found to improve.

INTRODUCTION

The ability of police departments to function efficiently is determined to a vast degree by the capability of officers to process information effectively (Russell, 1993). This procedure begins with the initial documentation of a crime scene and continues until disposed by the court. The documentation originates in the form of a police report which contains detailed information about witnesses, suspects, and recovered evidence. Accuracy and thoroughness are essential in a well written police report (Russell, 1993). However, many police jurisdictions have indicated that officers are unable to write effective reports (i.e., Wilson & Hays, 1984; McGough 1986; Johnson, 1987; Seay, 1988; Bauer & Shlechter, 1990; Kelly, 1990; Reardon, 1993; Russell, 1993; Pettaway, 1994).

Inefficiencies in Police Report Writing

Johnson (1987) contends that inefficiencies in police writing skills are interrelated involving aspects of composition & grammar, clarity, critical issues, vague or inaccurate statements of events by victims and witnesses, and organization. Wilson & Hays (1984) suggest that poor writing skills are rooted in an inability to interview, identify and process investigative information and produce accurate documentation. According to Seay (1988), there are often complaints of reports containing poor vocabulary, grammar and punctuation, inconsistencies in the reporting of events, and inaccuracies in describing evidence. Seay (1988) also noted that reports often fail to supply critical information, sometimes failing to respond to each allegation in an incident.

Consequences of Poor Writing Skills

There are several problems that may arise due to ineffectively written police reports. For instance, an accurate and thorough report is imperative in withstanding court challenges (Russell, 1993). Poor organization and grammatical errors could yield an unintelligible document (Seay, 1988), making its validity in court questionable. If a report fails to supply critical information, it may also be vulnerable to courtroom tests. When a report is unclear or poorly written it can elicit inaccurate assumptions, rendering false conclusions by the reader (Bauer & Shlechter, 1990). Not only is a clear and accurate report important in withstanding litigation, it is also necessary for aiding detectives in investigative procedures and counsel in preparation for prosecution (Russell, 1993).

Another problem involves time efficiency. A large portion of officers total available time is occupied by some type of information processing activity (Wilson & Hays, 1984). It has been estimated that officers spend 15-20% of their time engaged in report writing alone (Johnson, 1987; Russell, 1993). When a report is not written effectively, it is returned to the officer for a re-write, rendering considerable loss of productivity and wasted time for both the writer and the supervisor (Bauer & Shlechter, 1990).

Assessing Written Police Reports

The importance of effective written communication skills within the police department has been well documented (Wilson & Hays, 1984; McGough, 1986; Johnson, 1987; Seay, 1988; Bauer & Shelter, 1990; Kelly, 1990; Reardon, 1993; Russell, 1993). Yet, it has been found that many police jurisdictions do not have adequate methods for

evaluating report writing skills (Pettaway, Truxillo, Sulzer, & Warren, 1994). Although an abundance of literature has been published in this area, very few contain supporting empirical evidence.

Pettaway (1994) developed an instrument to assess officers' police report writing skills by means of an empirical research design. Critical dimensions of police report writing were defined and an instrument was constructed to assist police trainers in measuring officers' writing skills. The New Orleans Police Department Written Communications Test (NWCT) was developed to measure police written communication skills of New Orleans police recruits. Two forms of the NWCT were developed, each containing four subtests, modeled after the Test of Written Language-2 subtests (TOWL 2; Hammill & Larsen, 1988). One direct subtest consisted of subjects watching a video involving a crime scene, writing a narrative about what they saw, and then being measured on such areas as grammar, sentence structure, accuracy, and recall skills. The other subtest involved an assessment of vocabulary, spelling and style using job-related stimuli. The tests were scored by trained police raters.

The reliability of the narrative portion of the NWCT was estimated to be low as measured by test-rest, parallel forms, internal consistency, intraclass, interscorer, and generalizability analyses, while the reliability of the other subtests were found to be psychometrically sound. The test was found to lack internal consistency. Pettaway (1994) concluded that the lack of reliability evidence limits the probability of accumulating meaningful validity evidence. It was contended that further research of the scales used in assessment needed to be conducted in order to improve reliability.

Objectives

The goal of this project was to improve the reliability of the narrative subtest of the NWCT. This included an investigation of previously written communication assessments, rating techniques, and rater training effects in contribution to the development of new scales for evaluating the NWCT. It was anticipated that the reliability would increase when the scales were designed in a way as to limit raters' subjectivity in scoring. This was attempted by using well trained raters to score the narrative portion of the NWCT using behaviorally anchored scales.

Assessment of Written Communication

The difficulty of assessing written communication skills has been thoroughly investigated (i.e., Quellmalz, 1981; Hammil & Larsen, 1988; Mann, 1988; Mather, 1989; Bauer & Shlechter, 1990). However, previous literature does not provide definitive guidance for accurately measuring writing skills. The two most common procedures presently used are referred to as direct and indirect assessments (Mather, 1989).

Direct writing assessments are subjective measurements of written essays (i.e., the narrative portion of the NWCT). Both composition and basic skills can be evaluated within a writing sample. The direct approach has high face validity and positive user attitudes. However, it requires subjective measurement, often resulting in rater disagreement yielding low reliability (Bauer & Shelter, 1990). Indeed, the direct portion of Pettaway's (1994) NWCT was found to have low interrater reliability.

The most commonly used types of direct writing approaches are analytic and holistic (Mann, 1988; Quellmalz, 1981; Mather, 1989). Analytic assessment provides

more information about the proficiency of writing. A specific scoring guide or scale is established for measurement of defined components of the writing sample (Quellmalz, 1981). Holistic assessment responds to the writing composition as a unit. It is a single judgment of the overall quality. Although a quick and costly means of measurement, holistic rating does not assess specific components of writing (Mann, 1988).

Indirect writing assessments are measures of writing skills which require little, if any, actual writing (Mann, 1988). They are used to measure rules of writing (i.e., spelling, grammar, and punctuation) by responding to stimuli most commonly in a multiple-choice format. Scoring is easily determined yielding high rater agreement and high reliability (Mather, 1989), as was the case with the indirect portions of the NWCT. Validity evidence for indirect tests for written communication assessment, however, has been inconsistent. Because assessment is dependent upon reading without the production of written work, it has been argued that indirect formats utilize different psychological processes than required by production tasks (Quellmalz, 1981).

Critics supporting the indirect approach have contended that indirect measures of assessment have been shown to be moderately to highly correlated with general writing proficiency (Breland & Gaynor, 1979; Spandel & Stiggins, 1980; Quellmalz, 1981). However, recent studies demonstrate considerably lower correlations between direct and indirect approaches (Quellmalz, Capell, & Chou, 1982). The overall research suggests that direct methods are superior to the less definitive approach of indirect methods when measuring writing skills.

The direct portion of the NWCT was the focus of this investigation because of the importance of adequately assessing writing skills. Moreover, as previous work suggests, the indirect portions of NWCT showed high reliabilities while the direct portion of the test showed low reliabilities.

Behaviorally Anchored Rating Scales

The appropriate means for rating the police recruits' performance on the direct portion of the NWCT is of key importance in the writing skill evaluation. Performance evaluations have classically been rated based upon "traits" such as in the traditional graphic rating scale. This method employs poorly defined performance dimensions and scales which can lead to ambiguity and lack of independent ratings between dimensions (Schwab, Heneman, & DeCotiis, 1975).

Smith & Kendall (1954) introduced a variation in rating scales based upon Flannagan's (1949) work with "critical incidents". Critical incidents are described as the critical behaviors executed when determining whether a performance is good, average, or poor (Landy, 1989). Smith & Kendall (1963) developed The Behaviorally Anchored Rating Scale (BARS) to evaluate performance consisting of unambiguous anchors based upon this idea. They surmised that anchors on rating scales should be developed as statements which are capable of discriminating between good and poor performances. These anchors are examples of desired behaviors developed by groups of experts. These experts, often referred to as Subject Matter Experts (SMEs), consist of incumbents who have extensive knowledge of the most important aspects of the job (Landy, 1989).

BARS have many advantages compared to other means of measurement and have been seen as superior to traditional rating scales. One of the most appealing aspects of BARS is its high face validity. The behaviors determined as critical are identified by workers and supervisors and presented in a job-related language rather than that of a personnel director or psychologist (Campbell, Dunnette, Arvey, & Hellervik, 1973; Landy, 1989). Also, meaningful samples of actual job performance yield stronger content validity when presenting examples of job behavior rather than using measurements serving as indicators of behavior (Fogli, Hulin, & Blood, 1971). Producing a higher face valid and content valid rating scale should yield more reliable ratings (Dunnette, 1966).

Another advantage is the minimization of subjectivity resulting in less susceptibility to the classical rating errors (Barrette, 1966; Campbell, Dunnette, Arvey, and Hellervik, 1973). Since the scales are defined in concrete examples of behaviors, the evaluator makes fewer inferences. By reducing rating errors, the instrument is believed to yield a higher interrater reliability .

It has also been found that BARS produces relatively independent multidimensional measures of performance which will result in lower halo effect, again resulting in higher reliability. When dimensions are vaguely defined, the rater tends to treat them as once concept, yielding the higher halo effect (Dunnette, 1966; Campbell, Dunnette, Arvey, & Hellervik, 1973). Furthermore, these well defined dimensions also offer additional information on an individual's performance over that of globally based judgments (Fogli, Hulin & Blood, 1971).

BARS were developed in this study to replace the graphic rating scales used in Pettaway's (1994) investigation. It was anticipated that generating behavioral anchors would improve rater agreement.

Despite the proclaimed advantages of BARS, some contradicting findings have challenged this hypothesis of superiority. In a comparative study of global and numerical based rating instruments vs. BARS, Borman and Vallon (1974) found BARS to result in greater leniency error. Yet in a similar study, Campbell, Dunnette, Arvey, & Hellervick (1973) found BARS to result in less leniency error. In assessing interrater reliability Campbell et al., (1973) found BARS to yield slightly lower reliability than other rating methods, however, Borman and Vallon (1974) and Williams and Seiler (1973) found BARS to yield higher interrater reliability. On examination of dimension independence, Campbell et al. (1973) found behaviorally anchored dimensions to result in lower intercorrelations than numerically anchored instruments. However, Arvey and Hoyle (1974) and Borman and Vallon (1974) found no differences between behaviorally anchored instruments vs. alternative methods when assessing intercorrelations.

Rater Training

Although researchers continue to disagree over the superiority of BARS, studies continually have shown BARS to be superior when combined with rater training. According to Kendall & Butcher (1982), rater training helps to ensure acceptable levels of interrater reliability. Bernardin & Walter (1977), Borman (1975) and Borman & Dunnette (1975) claim that behaviorally based rating instruments will prove to be superior when combined with formal training of raters in using the instrument. Latham & Wexley (1981)

feel that training appears to be the critical factor in reducing error and thus increasing reliability in utilizing all rating scales. Fay and Latham (1982) compared rater training effects of three different rating scales. It was found that all the rating scales benefited from rater training, however, behaviorally based rating instruments showed the greatest improvement of rating errors.

Early approaches to rater training focuses on the reduction of classic psychometric rating errors. This type of training, Rater Error Training (RET), involves the presentation of definitions, graphic illustrations and examples of rating errors in an attempt to teach raters of common rating errors such as halo, leniency, and central tendency (McIntyre, Smith, & Hassett, 1984; Hedge & Kavanagh, 1988). Most studies show this type of training to be helpful in reducing these errors when using behaviorally based instruments (Borman, 1975; Latham, Wexley, & Pursell, 1975; Bernardin & Walter, 1977; Borman, 1979; Ivancevich, 1979). However, further research has shown this reduction of error had a negative effect on the actual accuracy of ratings (Borman, 1975; Borman, 1979; Bernardin & Pence, 1980; Smith, Hassett, & McIntyre, 1982). It has been suggested that while training to avoid classical rating errors, the learning of a new response set in rating behavior is facilitated, causing this decrease in the accuracy of a rater's responses (Borman, 1975; Borman, 1979; Bernardin & Pence, 1980; Bernardin & Buckley, 1981; Smith, Hassett, & McIntyre, 1982).

More recent research has begun to look at accurately rating dimensions rather than avoiding rater behaviors that lead to psychometric errors. Rater Accuracy Training (RAT) is the enhancement of accurate rating through the discussion of the multidimensionality of

work performance, and the importance of making fair and accurate ratings based upon well defined stereotypes of effective and ineffective performance (Bernardin & Pence, 1980).

Bernardin and Buckley (1981) used this idea of rater accuracy to develop a rater training procedure, Frame of Reference Training (FOR), to help raters accurately rate dimensions based upon a common frame of reference. The aim of this procedure is to develop standards for effective ratings congruent with those of the expert raters in an attempt to standardize raters' perceptions of behaviors (Bernardin & Buckley, 1981). McIntyre, Smith and Hassett (1984) have described the core of FOR as being: a) a description of the work to be evaluated, b) practice and feedback with ratings, and c) rationales for the ratings attributed to behaviors by the expert raters.

Research studies analyzing the effectiveness of FOR have been positive. Bernardin and Buckley (1981) contend that FOR actually increases interrater agreement. In an empirical study by McIntyre, Smith and Hassett (1984), FOR was found to be superior to traditional rater training in improving accuracy. Furthermore, Athey and McIntyre (1987) have shown FOR to facilitate better learning of training information and improve accuracy with less halo. FOR has been demonstrated as a successful means of training in many investigations and was chosen to be used in this project as a model to develop rater training in attempt to improve the reliability of the NWCT direct assessment scales.

METHOD

Raters

A group of 10 graduate students from the University of Houston-Clear Lake participated in this study on a voluntary basis. They were recruited in their classes to rescore the narrative portion of the NWCT (Form A) used in Pettaway's (1994) study.

Scale Revisions

The dimensions assessed in this study were based upon those used in the NWCT (1994) which were derived from job analysis results. SMEs from the Houston Police Department, two report writing trainers, reviewed the original dimensions and made suggestions for omitting and combining dimensions. The resulting list of dimensions can be found in Table 1.

Behaviorally Anchored Rating Scales

The rating scales for each of the nine dimensions consisted of the scale definition and a 5-point continuum, with the scores of 5, 3 and 1 defined by specified behavioral incidents. Each anchor was developed to represent a score that illustrated poor, average or excellent performance in job-related terms. The anchors were developed based on SME input, police report writing training materials, and Burry and Quellmalz 's (1983) method of defining behavioral incidents in assessing writing skills.. The anchors are presented in Table 2.

Rater Training

All subjects participated in a FOR training prior to scoring the tests. Each session lasted approximately one and one-half hours and had one to three participants. A description of the NWCT was provided. Subjects watched the video portion of the direct writing skills

test from the NWCT and were given a written copy of the script for reference. The raters then scored a sample test and the ratings were discussed in small groups of 2-3 with the expert rater. The purpose of the feedback was to develop standards for ratings congruent with those of the expert rater in an attempt to standardize the subjects' perceptions of poor, average and excellent responses.

Procedure

Fifty of the direct writing skills portion of the NWCT were rescored by the 10 subjects using the newly developed BARS. Subjects were given 10 tests to score. In order to compute reliability estimates, each test was scored by two subjects.

RESULTS

Reliability

A person's score on a test is composed of a "true" score plus some unsystematic error of measurement (Thurstone, 1931). *True score* is defined as the average of the scores that would be obtained if the person took the test an infinite number of times. The true score can never be determined exactly but must be estimated from the person's observed score on the test.

Since the variance of true scores cannot be computed directly (Guion, 1965), reliability is estimated by analyzing the effects of variations in administration conditions and test content on examinee's scores. Each method of calculating reliability takes into account somewhat different conditions that may produce unsystematic changes in test scores, thereby affecting the amount of error variance.

To determine the internal consistency of the new scales, *coefficient alpha* was computed. The alpha was relatively high, .82, for the new direct scale of the NWCT. Pettaway (1994) found alphas of .01 and -.06 for the direct scales of Forms A and B, respectively.

An *intraclass* correlation coefficient [ICC (1,1); Shrout & Fleiss, 1979] was computed for the scales to estimate interrater reliability. ICC (1,1) is appropriate for assessing interrater reliability when each subject is rated by a different set of raters. With subject's NWCT being scored by two raters, the ICC (1,1) was .62. Pettaway (1994) found an ICC of .51 for the total scores of both versions of the NWCT.

Interscorer reliability was computed for the raters' judgments of the new scales. The interscorer or interrater reliability was computed by correlating the subjects' total scores assigned by the two raters. The obtained interscorer reliability was .73 ($N = 50$, $p <$ .0001). Table 2 shows the interscorer reliabilities by dimension. As the table shows, some of the dimensions had higher interrater agreement than others (e.g., logical flow in narrative, $r = .59$; including only facts in narrative, $r = .80$), indicating that some of the scales are more reliable than others. Pettaway (1994) did not report similar results in her dissertation, but did report generalizability theory results which indicated very low reliabilities for the direct subtest.

DISCUSSION

Police training and selection literature have demonstrated that the accurate assessment of police officers's writing abilities is essential in ensuring that officers can be trained to have the skills they need to perform their jobs effectively. Pettaway's (1994) police written communication instrument, designed to assess the writing skills of New Orleans' police recruits, yielded high reliability on the indirect portion of the instrument, yet a low reliability on the direct portion. Written communication research has found that in order to adequately assess the written communication domain, direct assessments should be used. Accordingly, since validity is dependent upon high reliability, it was the goal of this study to improve the reliability of the direct portion.

A FOR training program was developed in an attempt to calibrate raters' perceptions of effective, average, and ineffective behaviors. In addition, behaviorally anchored scales were developed with the help of SMEs in an attempt to minimize the subjectivity of raters' judgements. The results suggest that the reliability of a direct writing skills assessment can be improved by these approaches.

As measured by Coefficient Alpha, an internal consistency of .82 showed a dramatic improvement over that of Pettaway's (1994) coefficients of .01 and -.06 for the direct scales of Forms A and B, respectively. In this case, the high internal consistency suggests that the new direct scales are indeed measuring aspects of a single dimension, police written communication.

As measured by an intraclass correlation coefficient [ICC, (1,1)], the interrater reliabilty was .62 for two raters. This shows a modest improvement over Pettaway's (1994) interrater reliability estimate of .51 for both Forms A and B.

Interrater reliabilty was high when correlating the total scores of the two raters using the newly revised scales. However, when correlated by dimension, it was found that some of the dimensions produced more agreement than others. For example, the dimension "Includes Sufficient Detail in Narrative About Facts of the Case" had a reliability of .41, whereas the dimension "Provides Description of Individuals Involved in the Crime Scene" had a reliability of .75. The latter dimension was more narrow in scope, requiring less subjective assessment by the rater. The anchors provided examples of direct types of characteristics to accurately assess this single idea. The former dimension, however, was more global, requiring the rater to make a subjective assessment of a number of possible attributes. Therefore, while the overall agreement for the new scales was high, some dimensions proved easier to rate than others.

Limitations

This investigation is limited in several areas. First, as mentioned previously, reliability might have been affected by the type of raters used. The raters were graudate students who were given no incentives to effortfully make accurate responses. Furthermore, these students were unfamiliar with the requirements of an effectively written police report as compared to utilizing experienced supervisors or police recruit trainers.

Also, the rater training program may have been limited due to time constraints. Each rater practiced with only one sample test, receiving minimal feedback. The implementation of a more thorough program for raters unfamiliar with the police officer's job could increase the standardization of rater's perceptions of performance, further improving reliability.

The reliability might have been higher had the BARS used in this assessment been developed on a more formal basis. As Smith and Kendall (1963) intended, the development of BARS involves a rigorous process including: 1) definition of critical incidents by job holders or supervisors, 2) assignment of incidents to a dimension, retaining only those dimensions that are proven to be consistently agreed upon by job holders/supervisors, and 3) having the job holders/supervisors rate the incidents' effectiveness level. By having each dimension well defined by job holders/supervisors and categorized by effectiveness, one should yield an instrument with equally strong dimensions throughout the test. Indeed, the BARS developed in this study only included the first step in Smith and Kendall's (1963) design.

More importantly, in utilizing a multidimensional instrument, it is benefical to limit each measure to one aspect of the performance in question. For example, the dimension "Logical Sentence Structure" contains anchors which include run-on sentences, logical sentence structure, incomplete sentences, and smothness of flow all under this one dimension. It is important to be precise and assess one particular behavior per dimension. In future scale construction, each dimension should contain only one thought.

Another problem with some of the scale definitions was that they were defined ambiguously. For example, "Spelling" had as anchors: *no spelling errors, some spelling errors,* and *the report is filled with spelling errors.* Ambiguous words such as "some" leave too much subjectivity in assessment, allowing raters with more than one way to interpret the dimension's intention. This dimension's reliability (r=.54) might have been improved had clearer definitions been generated.

Future Research

Further resarch is necessary to clarify the effects of a training program in conjunction with behaviorally based rating scales on interrater reliability for writing skill assessments. It is not clear if the improvement in reliability was due to the implementation of one of these methods or a combination of the two methods. Another important area of future investigation should involve rater accuracy retention. It is likely that positive changes by means of rater training will extinguish if they are not intermittently reinforced no matter how well the intial training porogram is developed.

Although the results of this study were positive, it is obvious that there is room for improvement. A more rigorous procedure in scale construction and development of a stringent rater training program may be the means of obtaining a direct writing skills assessment with high reliability. Alternatively, using raters who are familiar with the domains being evaluated might also improve reliability.

REFERENCES

Athey, T. R., & McIntyre, R. M. (1987). Effects of rater training on rater accuracy: Levels-of-Processing Theory and Social Facilitation Theory perspectives. Journal of Applied Psychology, 72, 567-572.

Arvey, R. D., & Hoyle, J. C. (1974). A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. Journal of Applied Psychology, 59, 61-68.

Bauer, R. K. & Shlechter, T. M. (1990). Evaluation issues related to writing skills of college educated professionals. Washington, DC: U.S. Army Armor School.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. Academy of Management Review, 6, 205-212.

Bernardin, H. J., & Pence, E. C. (1980). The effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Bernadine, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 62, 64-69.

Borman, W. C. (1975). Effects of instructions to avoid halo error in reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 60, 556-560.

Borman, W. C.,& Dunnette, M. D. (1975). Behavior-based versus trait oriented

performance ratings: An empirical study. Journal of Applied Psychology, 60, 561-565.

Borman, W. C., & Vallon, W. R. (1974). A review of what can happen when behavioral expectation scales are developed in one setting and used in another. Journal of Applied Psychology, 59, 197-201.

Breland, H. M., & Gaynor, J. L. (1979). A comparison of direct and indirect assessments of writing skill. Journal of Educational Measurement, 16, 119-128.

Burry, J. & Quellmalz, E. S. (1983). Assessing students' writing skills: The CSE Expository and Narrative Rating Scales. Los Angeles, CA: California University.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L.V. (1973). The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 57, 15-22.

Dunnette, M. D. (1966). Personnel Selection and Placement. Belmont: Wadsworth.

Fay, C. H., & Latham, G. P. (1982). Effects of training and rating scales on rating errors. Personnel Psychology, 35, 105-116.

Flannagan, J. C. (1949). A new approach to evaluating personnel. Personnel, 26, 35-42.

Fogli, L., Hulin, C. L., & Blood, M. R. (1971). Development of first-level behavioral job criteria. Psychological Bulletin, 55, 3-8.

Hammill, D. B., & Larsen, S. C. (1988). Test of Written Language-2. Austin, TX: PRO-ED.

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance

evaluations: Comparison of three methods of performance appraiser training. Journal of Applied Psychology, 73, 68-73.

Ivancevich, J. M. (1979). A longitudinal study of the effects of rater training on psychometric errors in ratings. Journal of Applied Psychology, 64, 502-508

Johnson, D. M. (1987). Attacking the report writing problem. The Police Chief, 54, 22-26.

Kelly, P. T. (1990). Increasing productivity by taping reports Police Chief, The 57, 49-50.

Kendall, P. C., & Butcher, J. N. (1982). Handbook of Research Methods in Clinical Psychology. New York: Wiley.

Landy, F. J. (1989). Psychology of Work Behavior (4th ed.). California: Brooks/Cole.

Latham, G. P., & Wexley, K. N. (1981). Increasing productivity through performance appraisal. Reading, MA: Addison-Wesley.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 60, 550-555.

Mann, Rebecca (1988). Measuring writing competency. Paper presented at the Southeastern Conference on English in the Two-Year college, Louisville, KY.

Mather, N. (1989). Comparison of the new and existing Woodcock-Johnson Writing Tests to other writing measures. Learning Disabilities Focus, 4, 84-95.

McGough, M. O. (1986). Writing police reports with portable computers. The

Police Chief, 53, 95.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 60,556-560.

Pettaway, M. (1994). The Development and validation of the New Orleans police recruit written communications test. Unpublished doctoral dissertation, Tulane University, New Orleans.

Quellmalz, E. S. (1986). Writing skills assessment. In R. Berk (Ed.), *Performance Assessment: Methods and Applications* (pp. 492-508). Maryland: Johns Hopkins University Press.

Russell, M. J. (1993, September). Toward the paperless police department: The use of laptop computers. *National Institute of Justice. Washington, DC:*.

Schwab, D. D., Heneman, H. G., & Decotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 28, 549-562.

Seay, W. T. (1988). Report writing, do it right the first time! FBI Law Enforcement Bulletin, 57, 2-4.

Smith, D. E., Hassett, C. E., & McIntyre, R. M. (1982, April). Using student ratings for administrative decisions: Are ratings contaminated by perceived uses of the information. Paper presented at the 23rd annual meeting of the Western Academy of Management, Colorado Springs, CO.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of

Applied Psychology, 47, 149-155.

Wilson, J. B. & Hayes, S. P. (1984). The competence-based approach to report writing: a new option. The Police Chief, 51, 26-29.

APPENDIX A

Definitions of Behaviorally Based Anchors

Logical flow in narrative

5 - The report is written in the same sequence in which the events occurred in the video. There is a logical and accurate flow in the written statement of events.

3 - The report is somewhat logical in flow, yet there is occasional incorrect sequencing of events. 1 - The report is scrambled and difficult to follow. There is no logical sequence.

Includes sufficient detail in narrative about facts of case

5 - The report contains all pertinent information regarding the crime scene. Entails names, times, dates, and location. Description of evidence, type of crime scene, and facts of the altercation are included and accurate.

3 - Some of the pertinent information is missing or is not relayed accurately.

1 - Most of pertinent information is either not included or relayed inaccurately.

Includes only facts in narrative

5 - The report includes only facts presented from the video. The examinee makes no presumptions and does not include any extraneous information.

3 - Facts are presented, however, there is some irrelevant or fabricated information included in the report.

1 - The report is filled with superfluous information. The examine includes presumptions and/or irrelevant information.

Provides description of individuals involved in the crime scene

5 - Accurately includes all important characteristics of each individual involved, including name, race, sex, apparel, and physical characteristics.

3 - Traits and characteristics are included, but report is missing some important details in describing some of the individuals involved or reports details inaccurately.

1 - Includes few, if any, details in describing individuals involved in the crime scene.

Vocabulary

5 - Proper use of all vocabulary words in sentences or context.

3 - At times there is misuse of a vocabulary word, rendering the report occasionally confusing.

1 - Constant misuse of vocabulary words. The report is difficult to understand.

Spelling

5 - Names, addresses, and important words are all spelled correctly.

3 - There are some spelling mistakes, however, the report is still comprehensible.

1 - The report is filled with spelling errors.

Punctuation

5 - Punctuation is accurate, enhancing understanding of sentences.

3 - Has a good grasp of basic punctuation rules, though occasional mistakes make some sentences confusing.

1 - Uses no punctuation or continually uses punctuation incorrectly.

Grammar

5 - Knows basic grammatical skills of writing. Has subject/verb agreement. Uses correct tenses in describing events.

3 - Has a good grasp of basic rules, though there are some grammatical errors.

1 - Little, if any, basic grammatical skills. Continual improper tenses used, little subject/verb agreement. The report is difficult to follow.

Logical sentence structure

5 - Combines short clauses into logical complete sentences. Avoids run-ons. The report reads smoothly.

3 - Adequate sentence structure, although there are occasional run-ons or incomplete sentences.

1 - Frequent incomplete sentences and/or run-ons. The report lacks logical sentence structure and is difficult to follow.

Table 1

List of Writing Skill Dimensions

Logical Flow in Narrative
Includes Sufficient Detail in Narrative About Facts of the Case
Includes only facts in narrative
Provides description of individuals in the crime scene
Vocabulary
Spelling
Punctuation
Grammar
Logical Sentence Structure

Table 2

Interrater reliabilities by dimension

DIMENSION	RELIABILITY COEFFICIENT
Logical Flow in Narrative	.58923
Includes Sufficient Detail in Narrative About Facts of the Case	.40782
Includes only facts in narrative	.79870
Provides description of individuals in the crime scene	.75154
Vocabulary	.61757
Spelling	.54529
Punctuation	.61647
Grammar	.60761
Logical Sentence Structure	.64979